Московский Государственный Технический Университет

им. Н.Э. Баумана



Отчет по лабораторной работе №1

по курсу

Технологии Машинного Обучения

# Выполнила:

## Костян Алина

## ИУ5-53

\_\_\_\_\_\_\_\_\_\_\_\_\_

# Проверил:

## Гапанюк Ю.Е.

\_\_\_\_\_\_\_\_\_\_\_\_\_

Москва, 2019

# Задание

* Выбрать набор данных (датасет).
* Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

* Сформировать отчет и разместить его в своем репозитории на github.

# Код и результаты выполнения

1. Выберем набор данных:

Этот набор данных предназначен для прогнозирования перспективы приема студентов из Индии.

Набор данных содержит несколько параметров, которые считаются важными при подаче заявки на магистерские программы. Параметры включают в себя:

1. GRE баллов (из 340)
2. TOEFL баллов (из 120)
3. Университетский рейтинг (из 5)
4. Заявление о цели и рекомендательное письмо сила (из 5)
5. Бакалавриат Средний балл (из 10)
6. Опыт исследования (0 или 1)
7. Вероятность поступления (от 0 до 1)
8. Подключаем библиотеки:

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import numpy as np

1. Основные характеристика датасета

Первые 5 строк (рис 1)

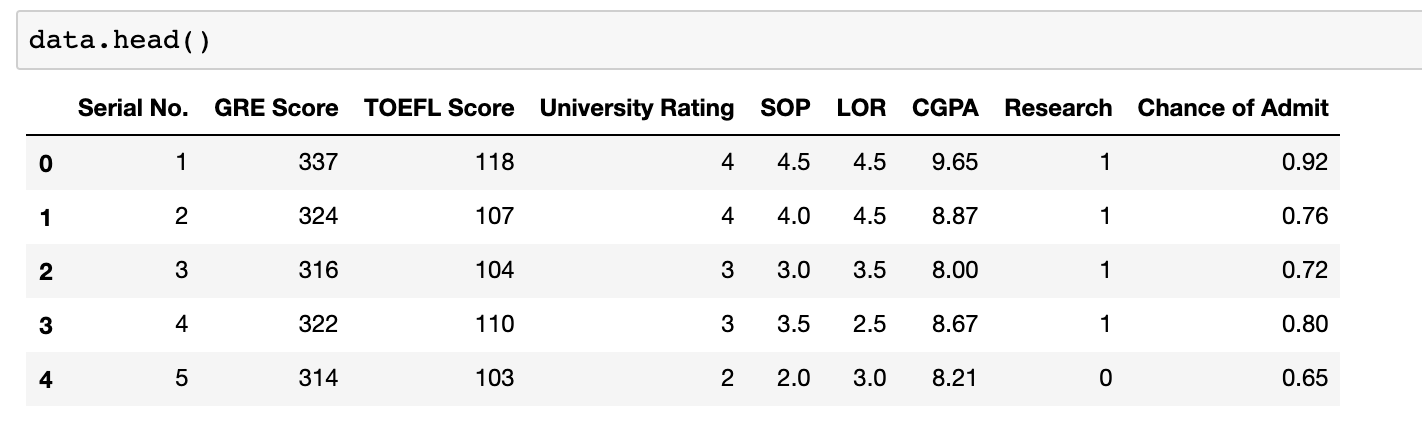


Рис. 1

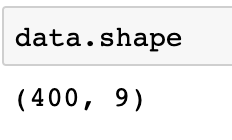
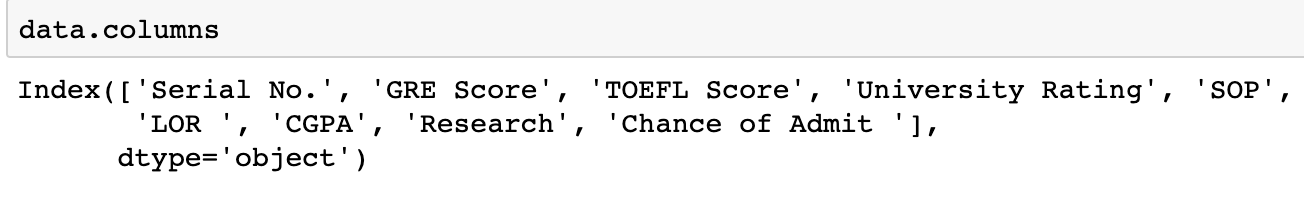
 

Рис. 3

Рис. 2

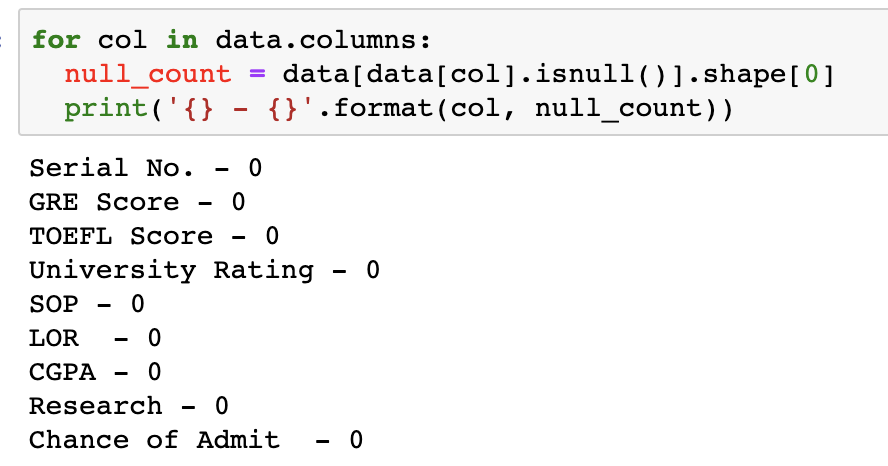
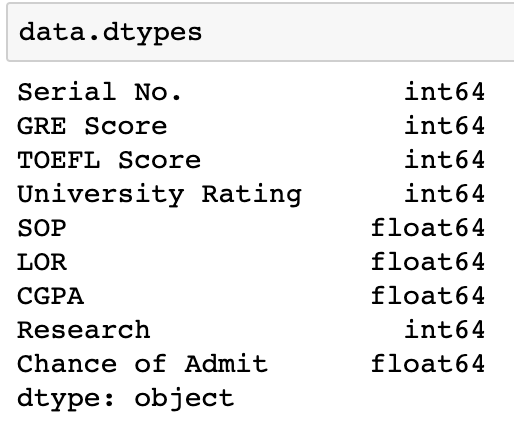


Рис. 4

Рис. 5

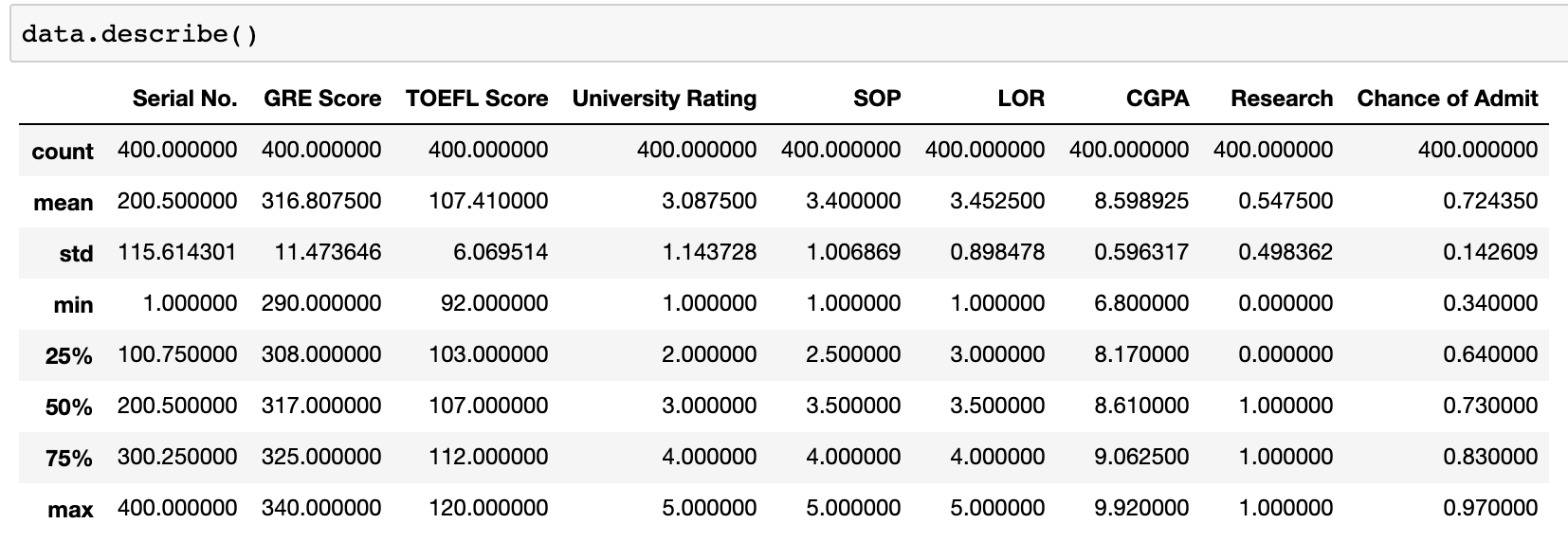


Рис. 6

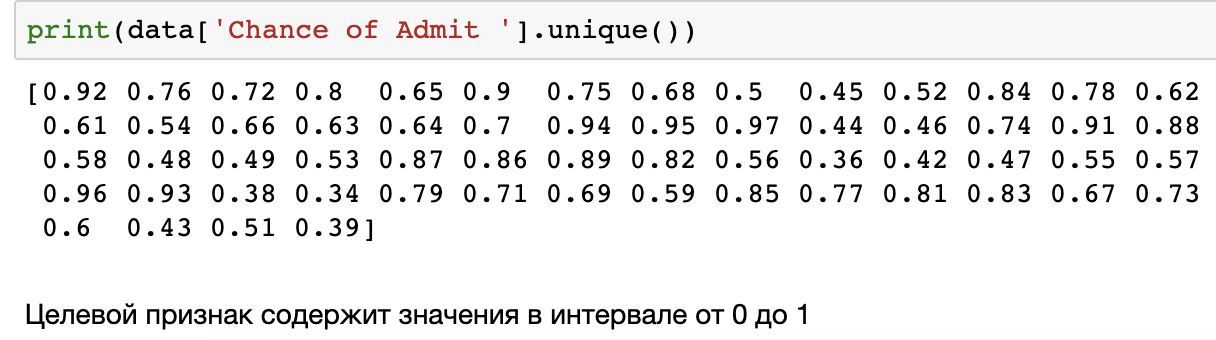


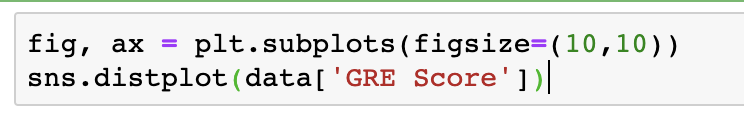
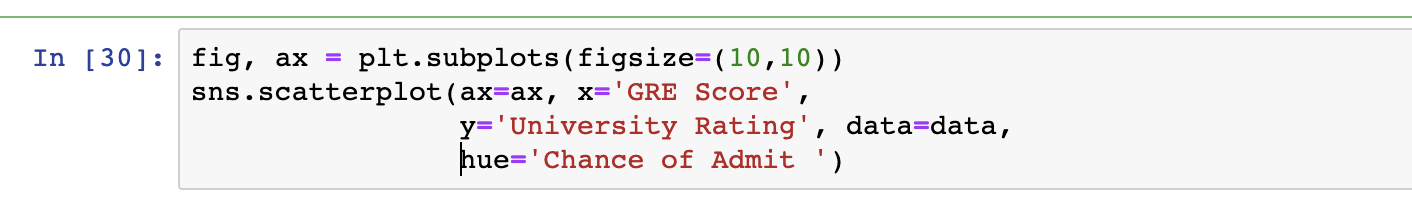
Рис. 7

1. Визуальное исследование набора данных
   1. Диаграмма рассеивания и влияние на нее целевого признака (рис. 8 и рис. 9)

Можно видеть, что несмотря на разброс значений, между количеством GRE баллов и Рейтингом университета есть почти что линейная зависимость. Чем лчше университет, тем больше баллов получит студент.

Видно, что шанс поступления гораздо выше, если у ВУЗа хороший рейтинг и хорошо сдан GRE.

* 1. Гистограмма – плотность распределения данных (рис. 10 и рис. 11)



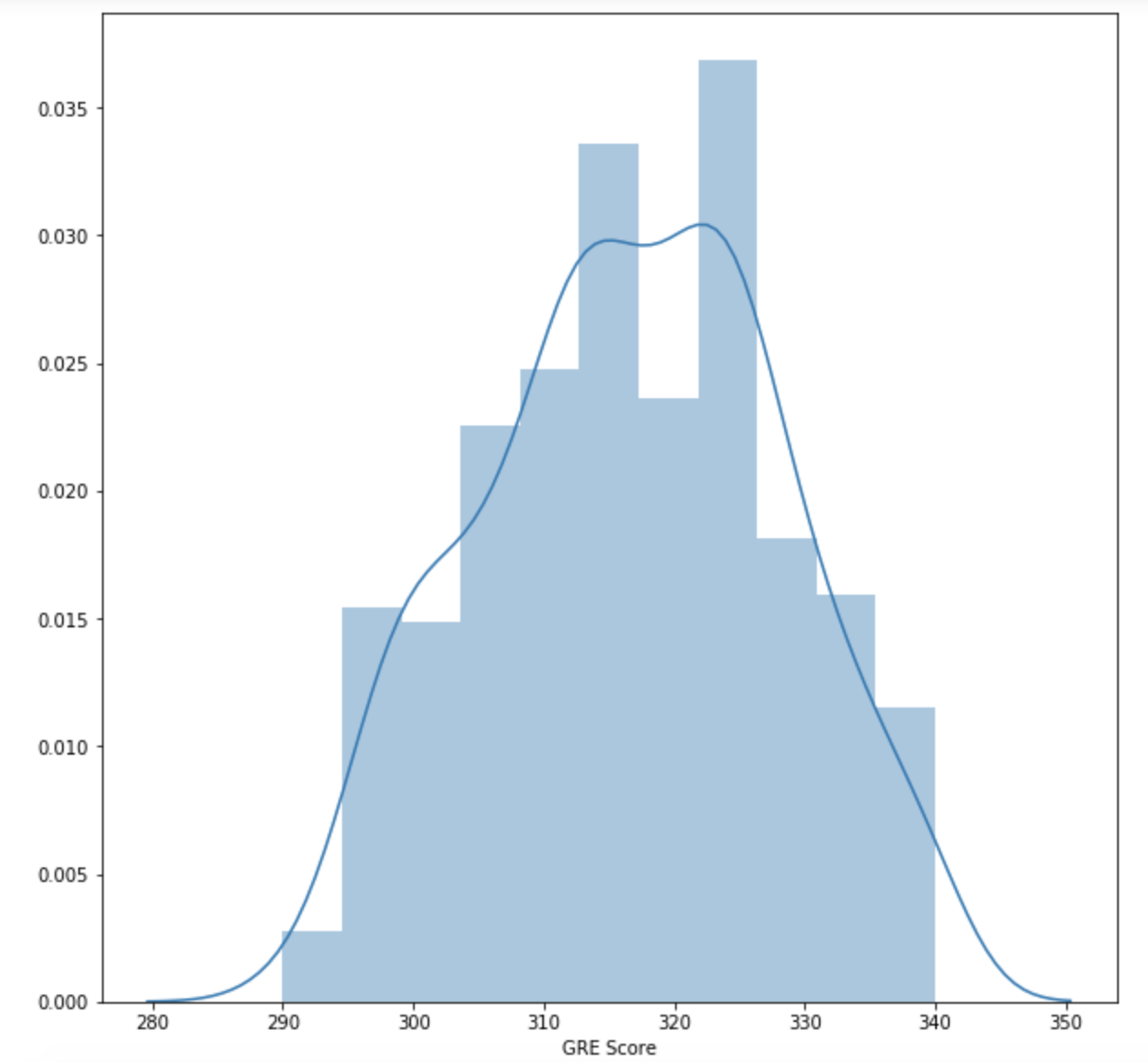
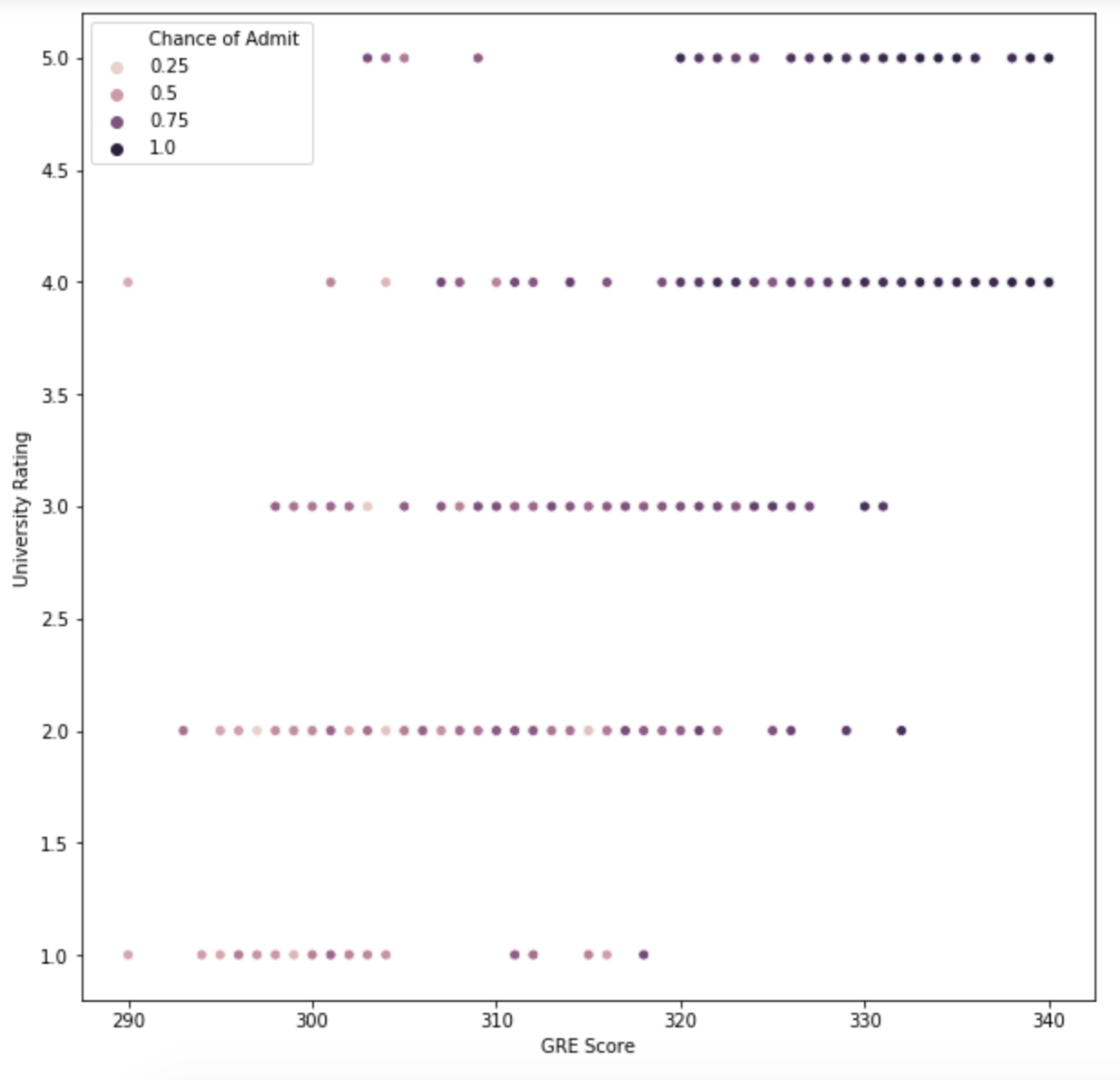


Рис. 10 и Рис. 11

Рис. 8 и Рис. 9

* 1. Joinplot - Комбинация гистограмм и диаграмм рассеивания. (рис. 12, рис. 13 и рис. 14)



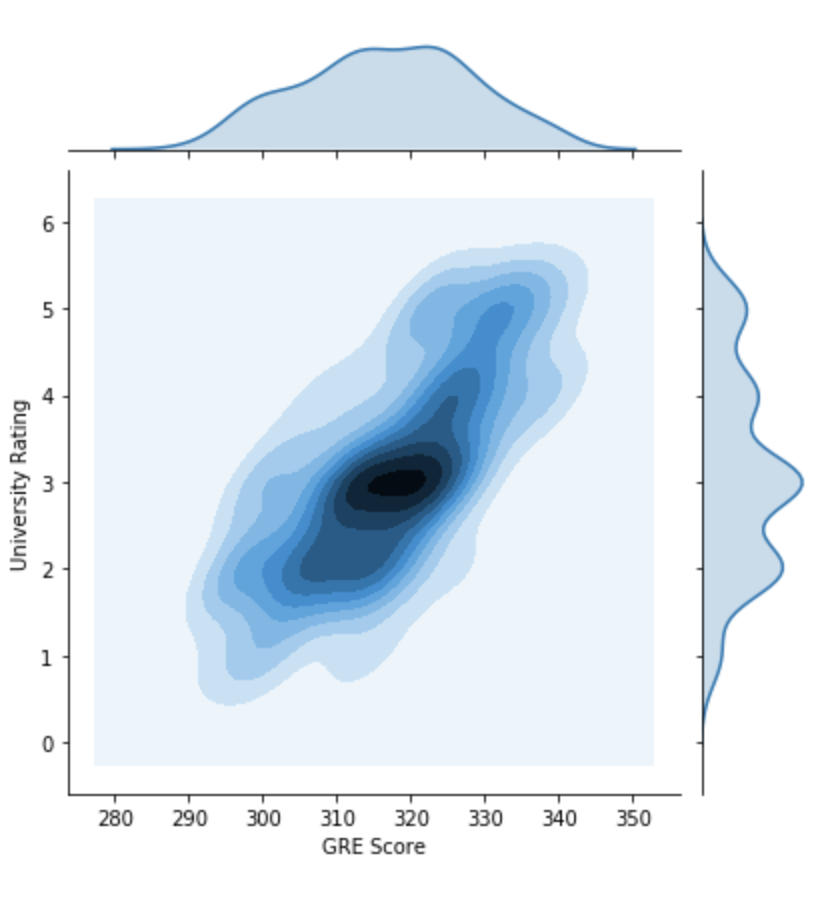
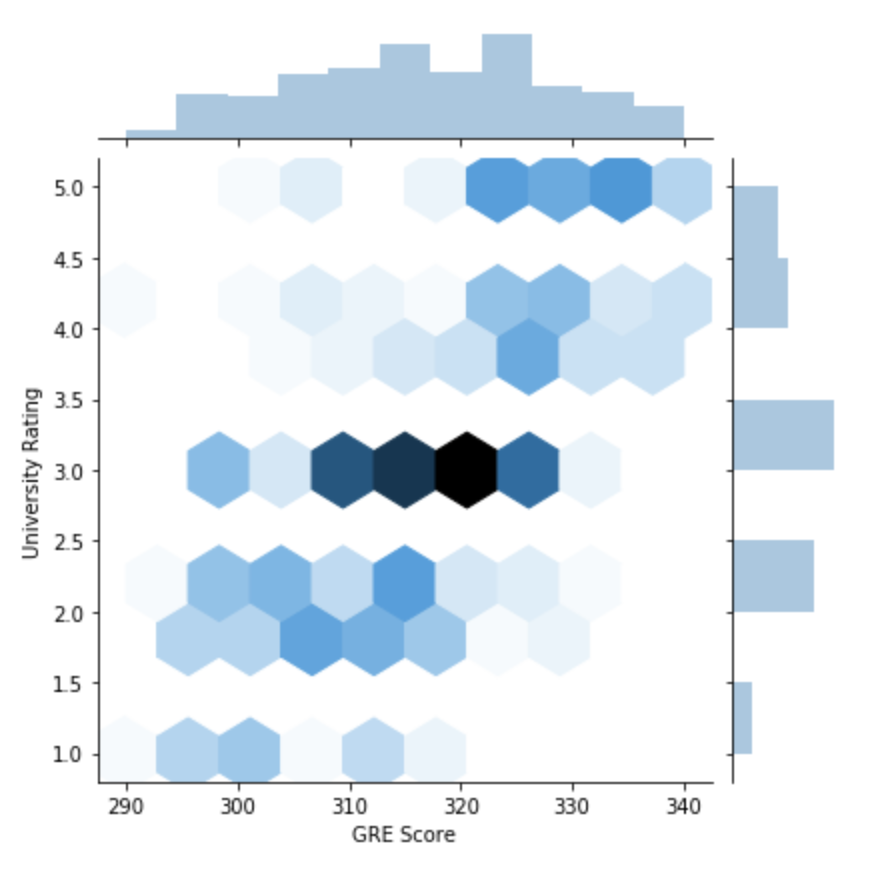
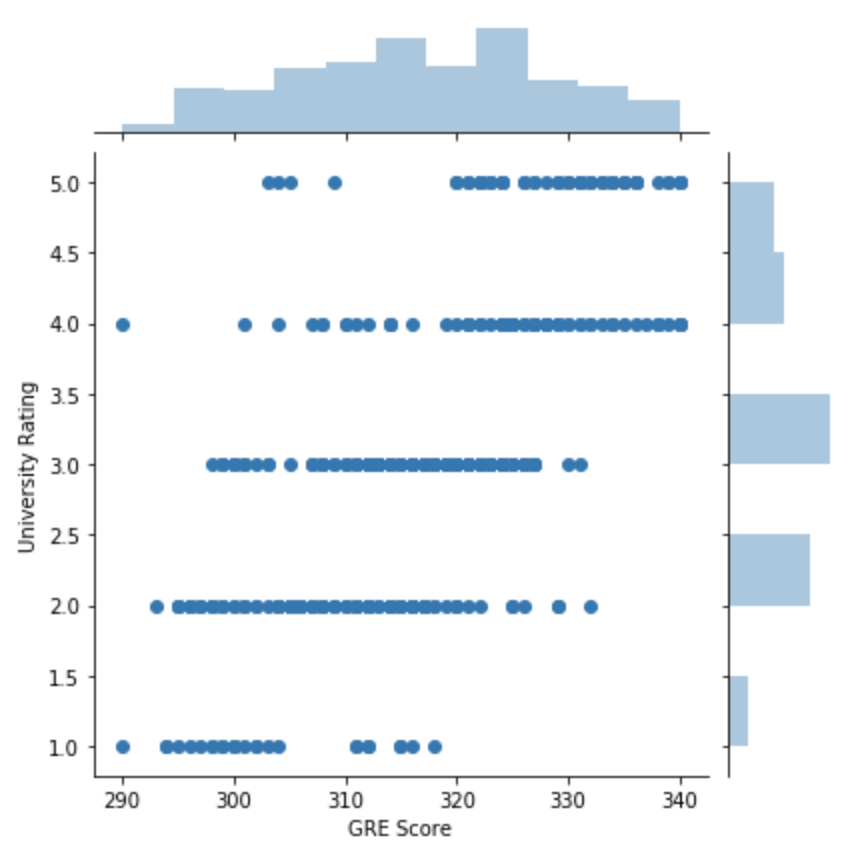


Рис. 12, Рис. 13 и Рис. 14

* 1. Ящик с усами - Отображает одномерное распределение вероятности. (рис. 15 и рис. 16)

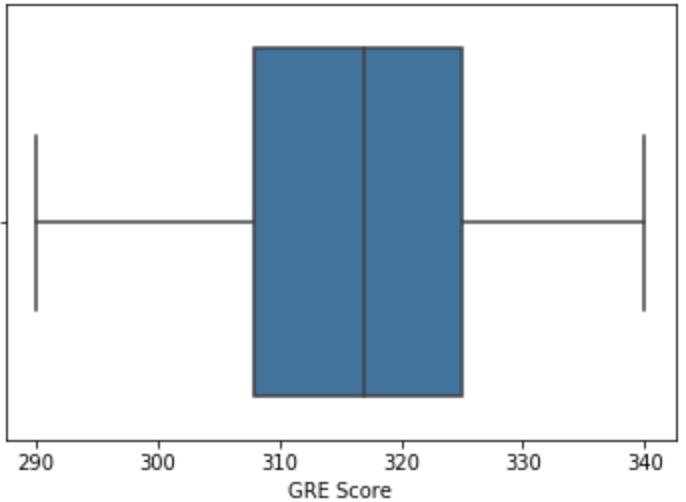


Рис. 15 и Рис. 16

* 1. Violin Plot (рис. 17)

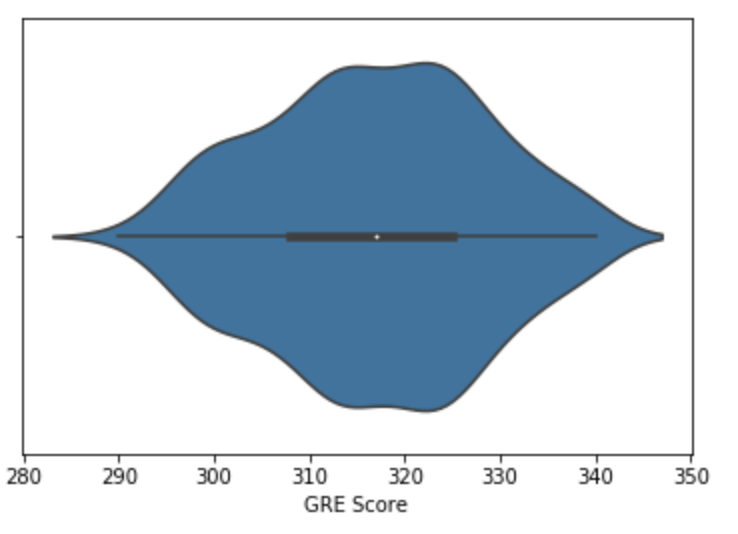


Рис. 17

* 1. Парные диаграммы для всего набора данных (рис. 18 и рис. 19)



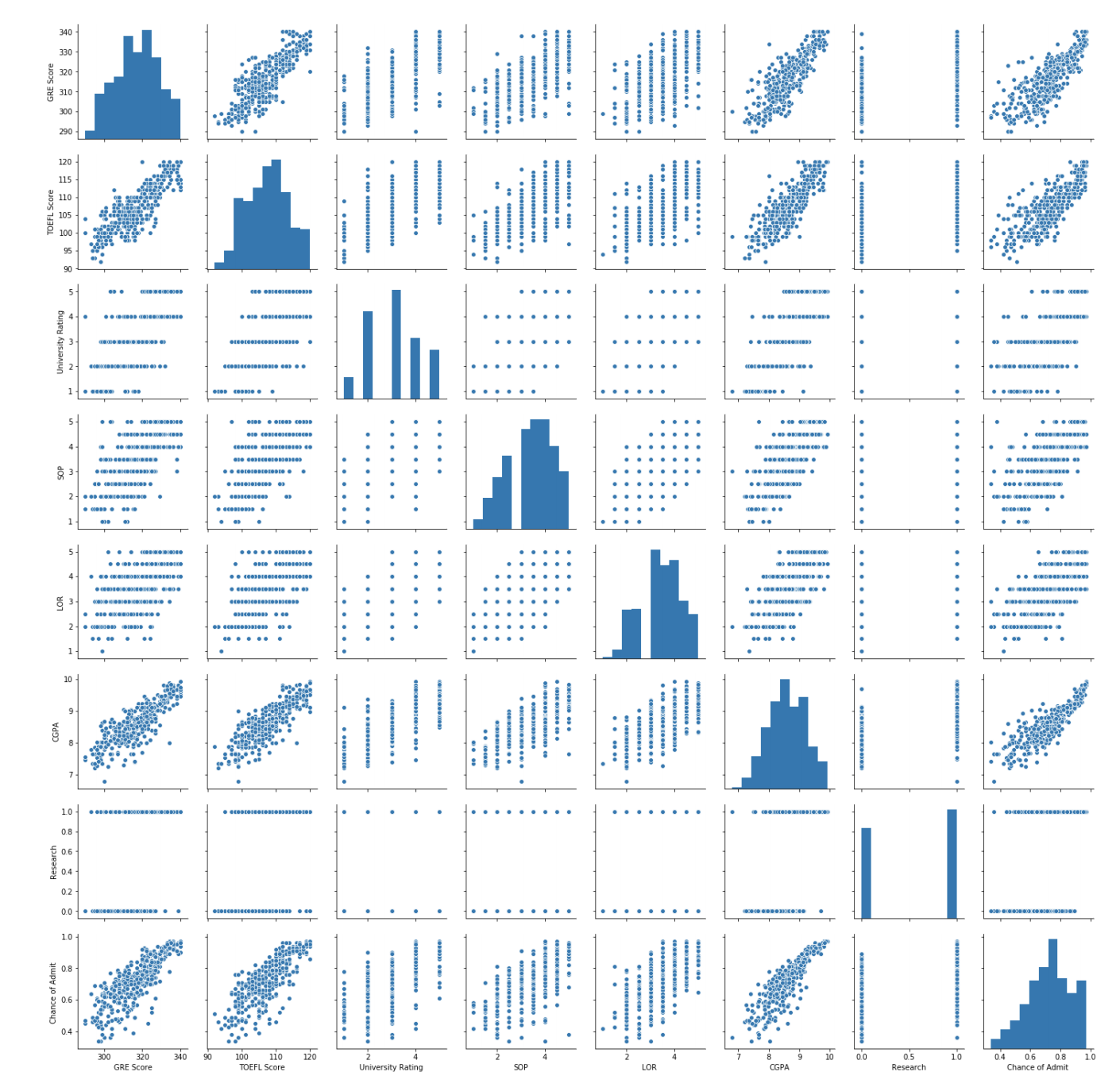


Рис. 18 и Рис. 19

1. Информация о корреляции признаков (рис. 20)

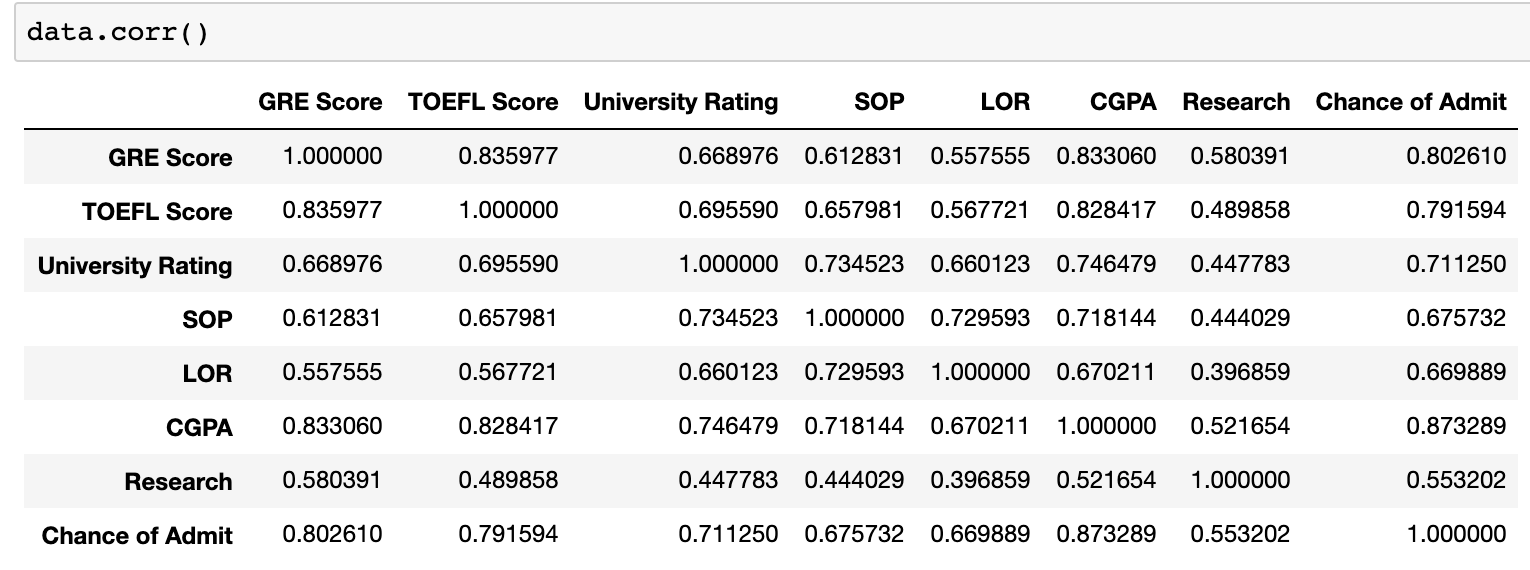
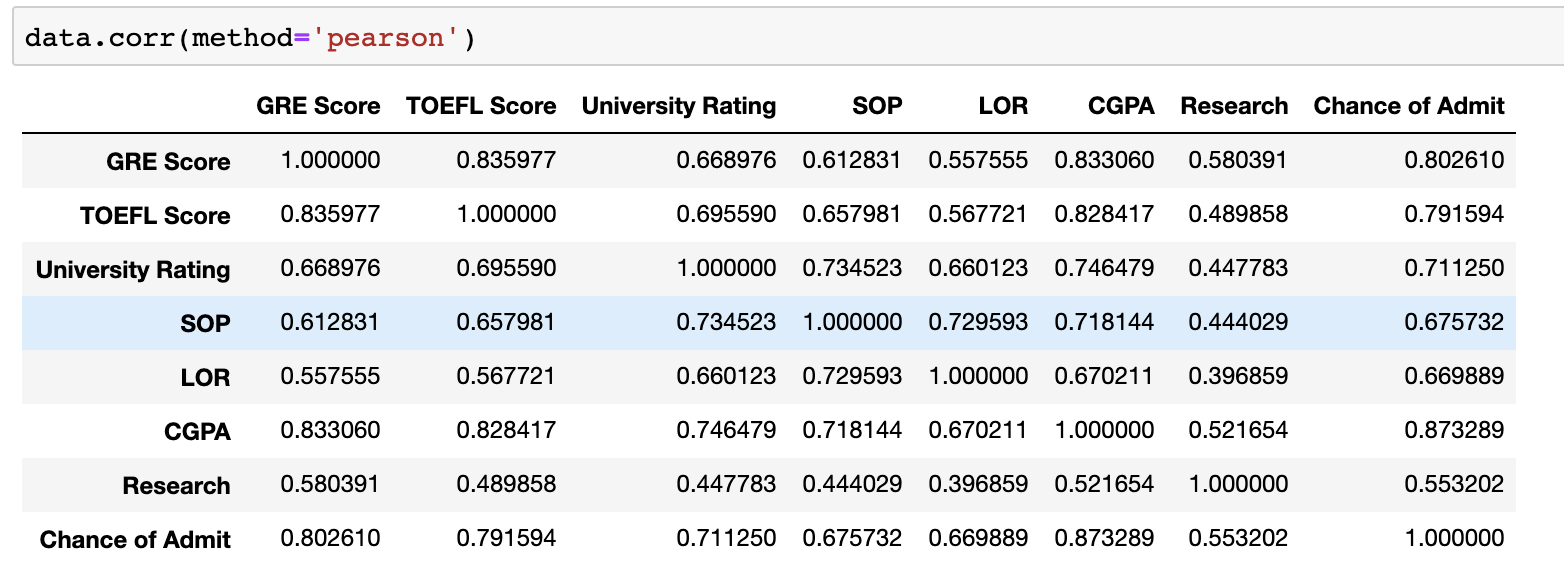
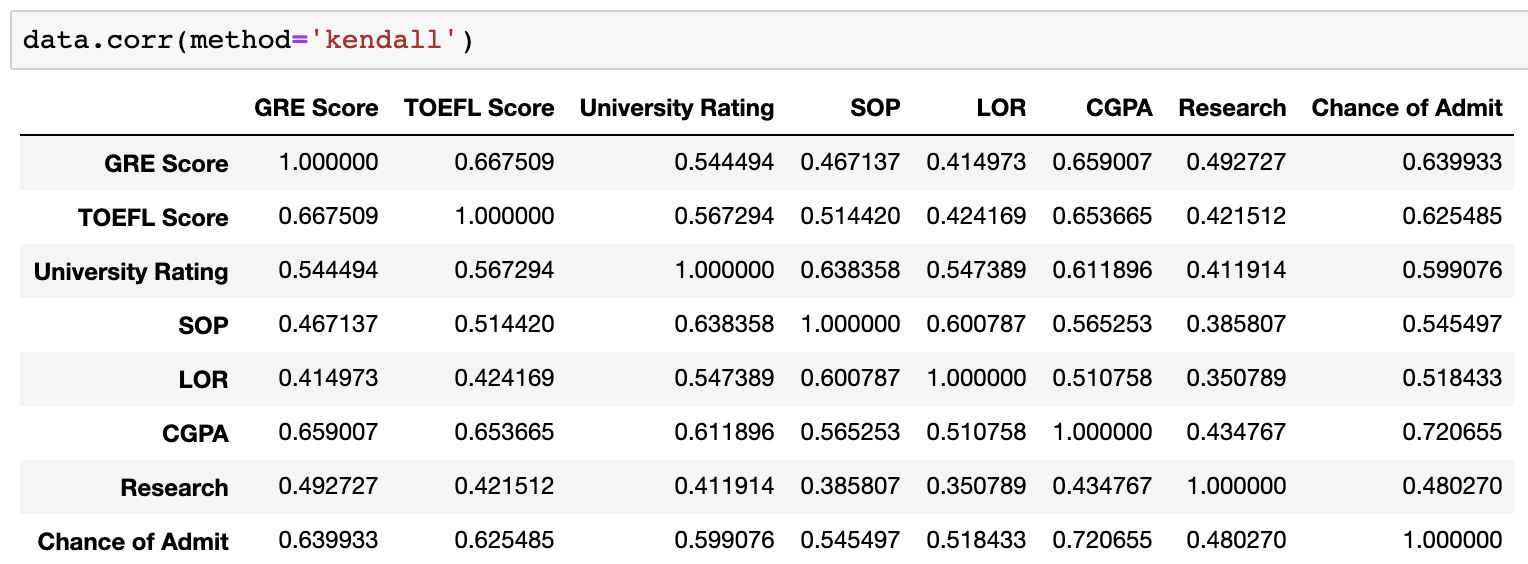


Рис. 20

Целевой признак наиболее сильно коррелирует с CGPA (0.87) и GRE (0.8). Эти признаки обязательно следует оставить в модели. Целевой признак отчасти коррелирует со всеми признаками из них нечего удалить.

Таблицы корреляции, заполненные разными способами. (рис. 21, рис. 22 и рис. 23)

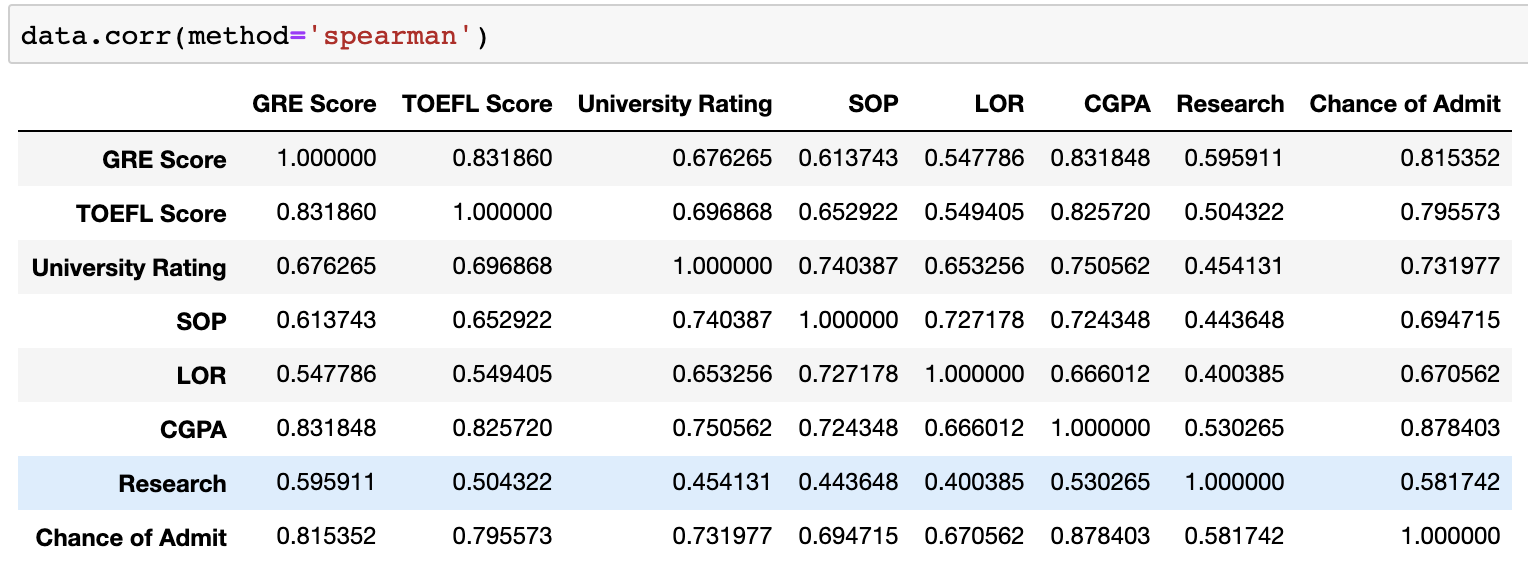


Рис. 21, Рис. 22 и Рис. 23

А так же корреляционные матрицы. (рис. 24)

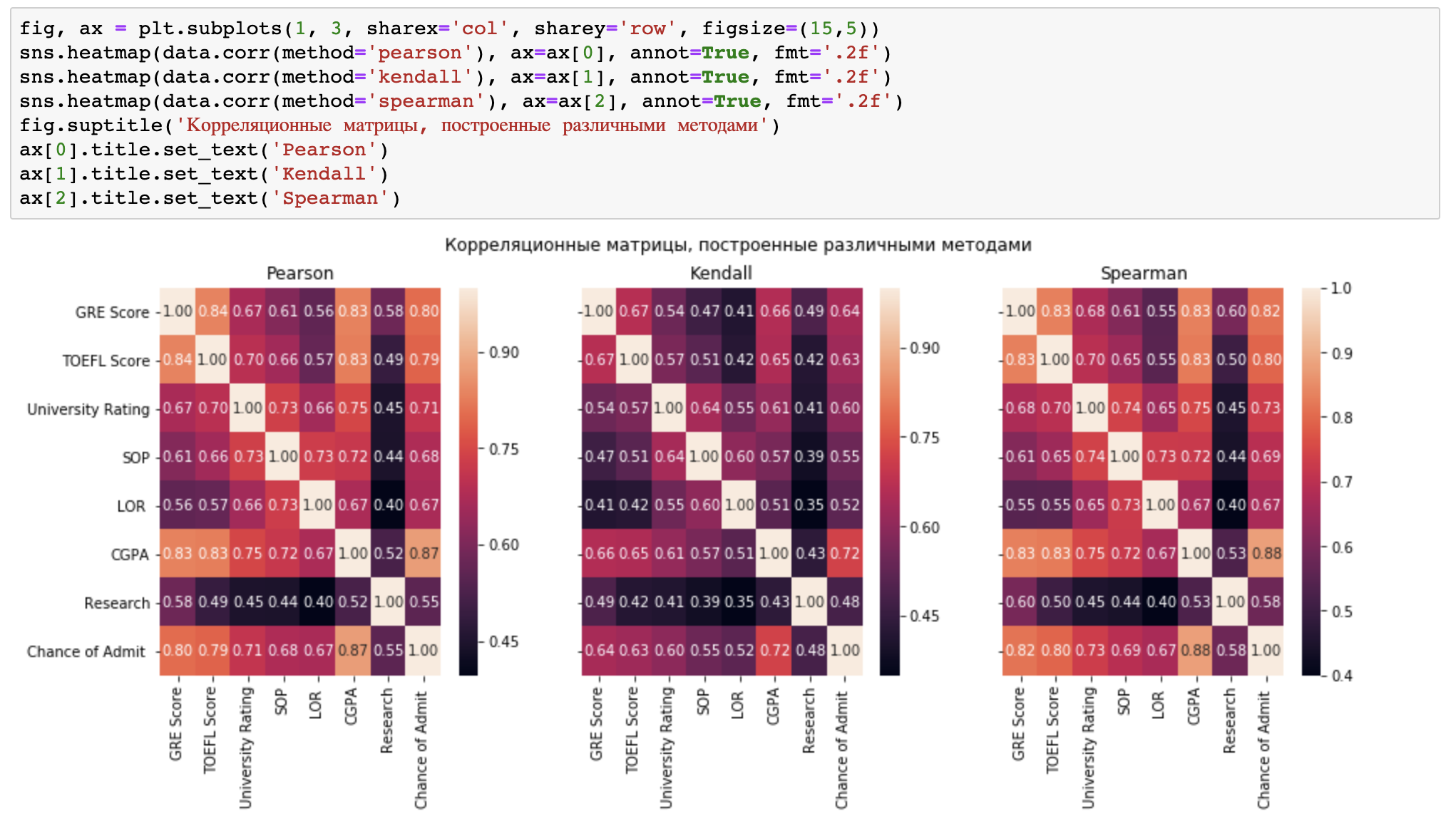


Рис. 24